Confronting the elephant in the room: Cleaning and Wrangling Data for Collections and Scholarly Services

Shannon Burke
Coordinator of Monograph Acquisitions

Clarke Iakovakis
Scholarly Services Librarian

View data, code, and other materials at https://osf.io/a5p3r/

Electronic Resources and Libraries

The Elephant in the Room

"Elephant" photograph courtesy Derek Hatfield on Flickr. Licensed under CC BY 2.0.

Data wrangling has long been an elephant in the room of data analysis.

Extraordinary amounts of time are spent getting a data set into a shape that is suitable for downstream analysis tools, often exceeding the amount of time spent on the analysis itself.

Sean Kandel et al., "Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data," *Information Visualization* 10, no. 4 (2011): 273.
Open Access version: http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

The Elephants In the Room

Data science skills

Need for clean & interoperable data

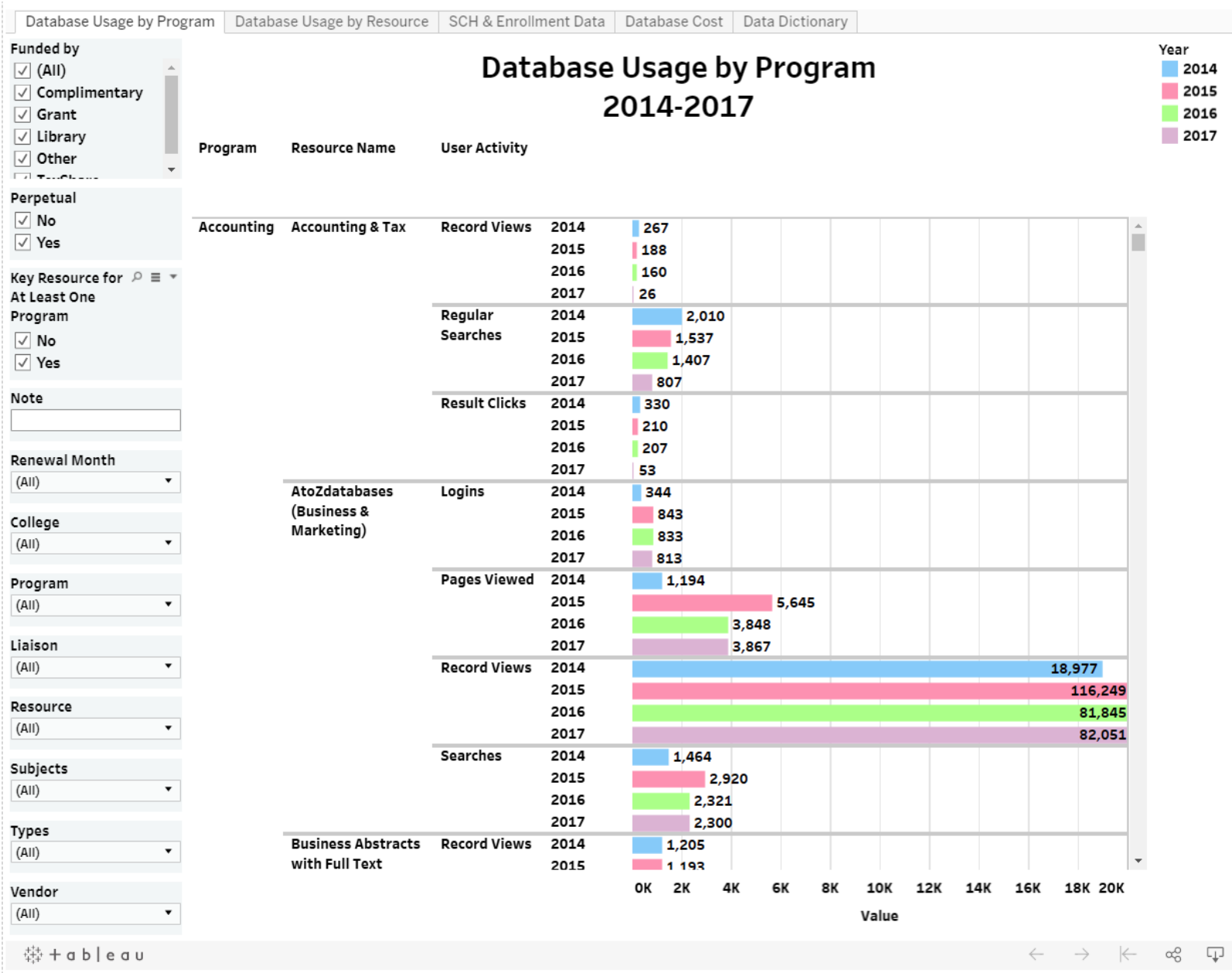
Administrative support

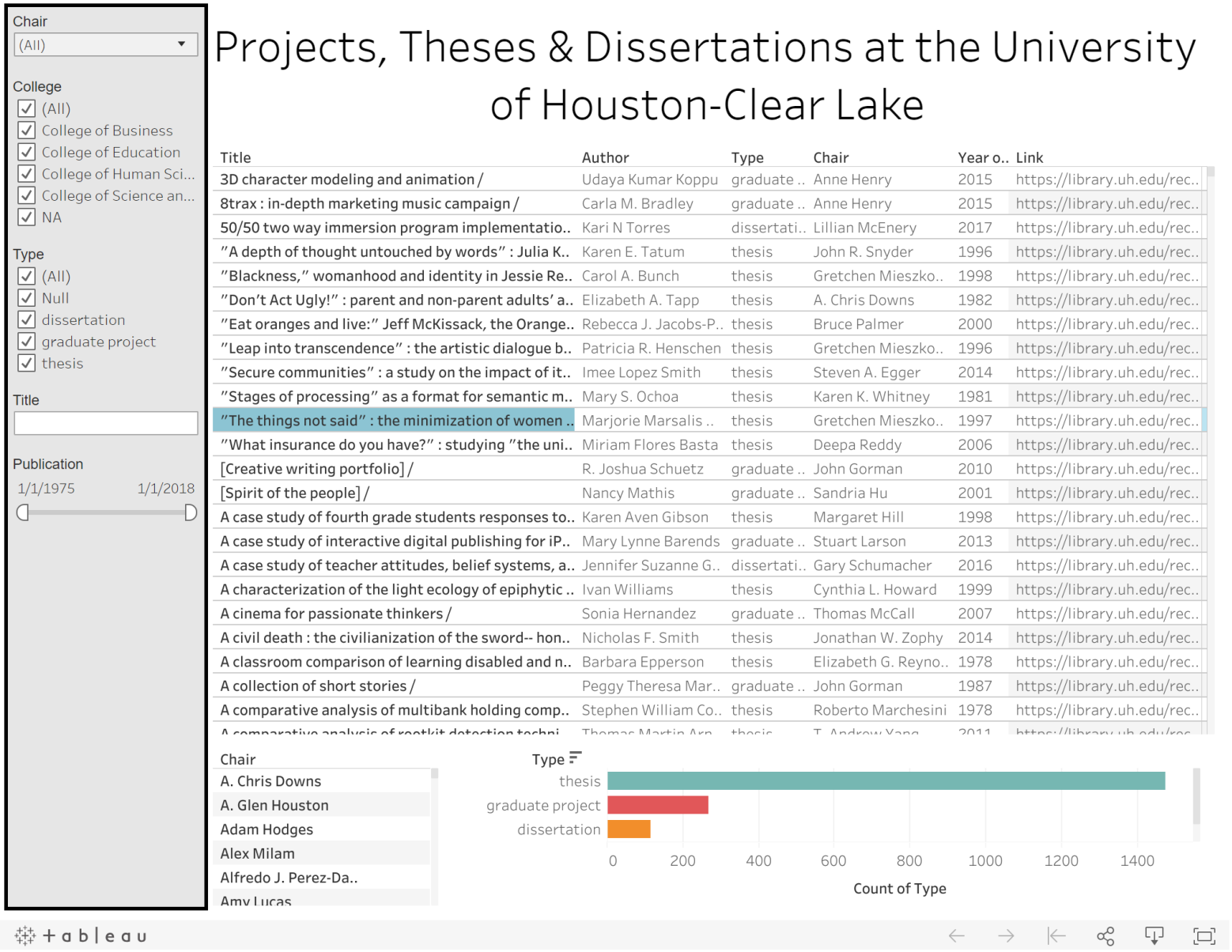
Data Culture

Reproducibility

Evaluating database usage over several years to make renewal/cancelation decisions

https://tabsoft.co/2C8dXpE



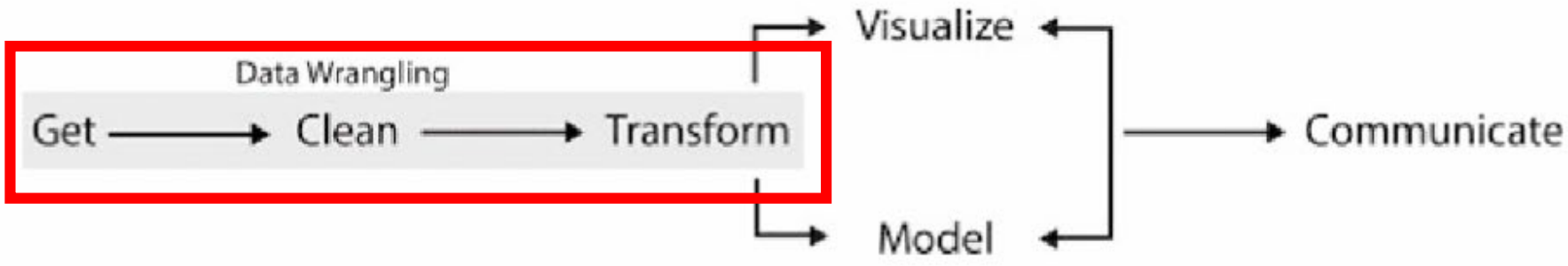


Complete thesis & dissertation submissions, with chairs, authors, titles, years, etc.

https://tabsoft.co/2C6BgQr

https://osf.io/a5p3r/

Data Wrangling

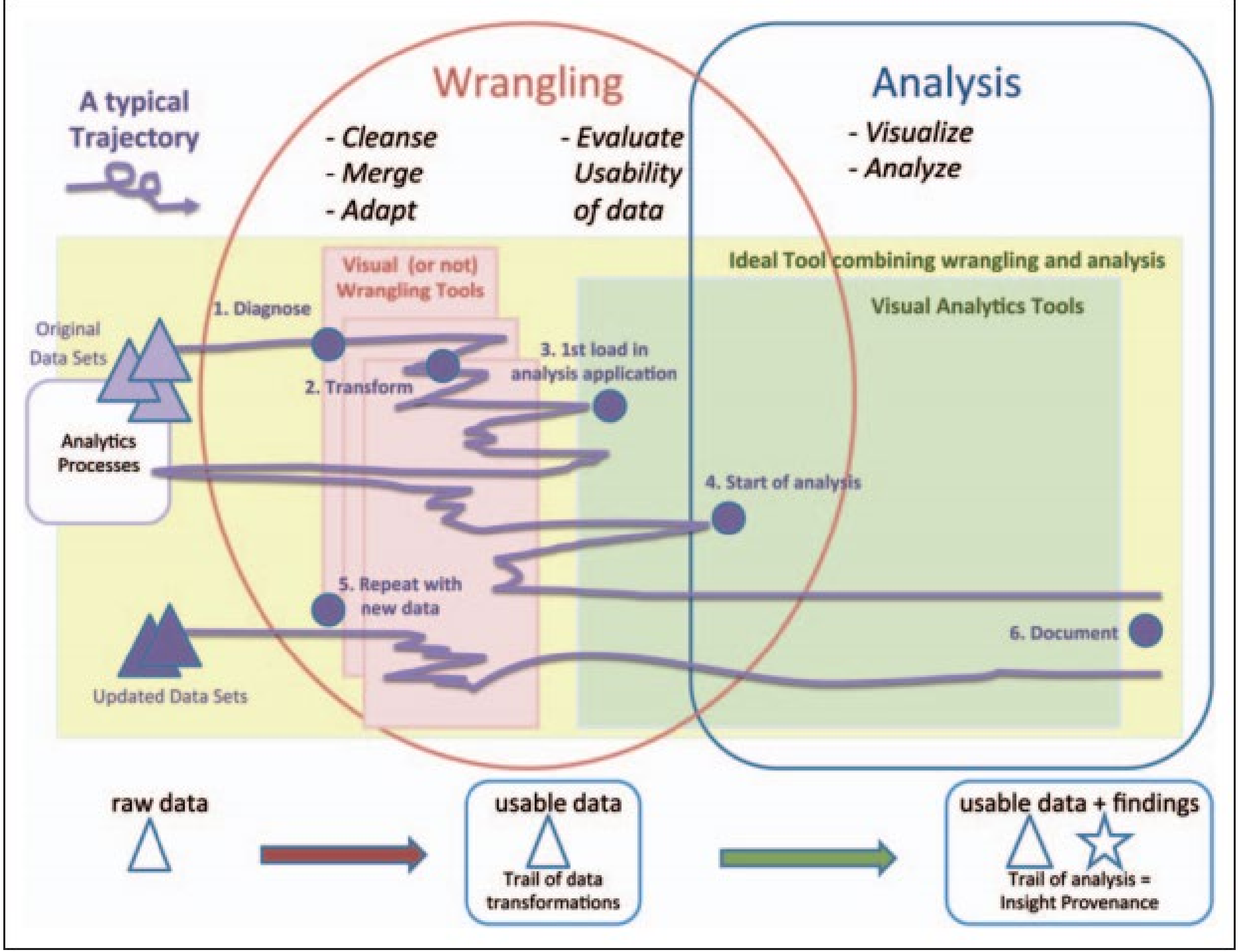


Bradley C. Boehmke, *Data Wrangling with R* (Switzerland: Springer, 2016), 4.

The ability to take a messy, unrefined source of data and wrangle it into something useful. The art of…extracting raw data and creating **clear and actionable bits of information** for your analysis.

Bradley C. Boehmke, *Data Wrangling with R* (Switzerland: Springer, 2016), 4.

A process of **iterative** data exploration & transformation that enables analysis.



An Absolute

Wrangling data is essential and cannot be ignored. Our questions are too complex and multifaceted to rely on a single, already clean and prepared dataset.

Wrangling is necessary to get to the analysis and visualization phase so that the data becomes actionable, enabling us to reach the decision-making stage.

Data Wrangling Challenges

Photograph courtesy **cskk** on Flickr at https://flic.kr/p/9Und1D. Licensed under CC BY-NC-ND 2.0.

Data Science in Libraries Project

Skills Gap

While practicing librarians are learning some data science skills, it is through ad-hoc, uncoordinated continuing education programs.

Management Gap

Library administrators need toolkits and frameworks to strategically use data science for data-driven decision making and management of library operations.

Burton, Matt and Lyon, Liz and Erdmann, Chris and Tijerina, Bonnie (2018) *Data Science in Libraries.* Presentation. University of Pittsburgh, Pittsburgh, PA. https://www.cni.org/wp-content/uploads/2017/12/cni_data_tijerina.pdf

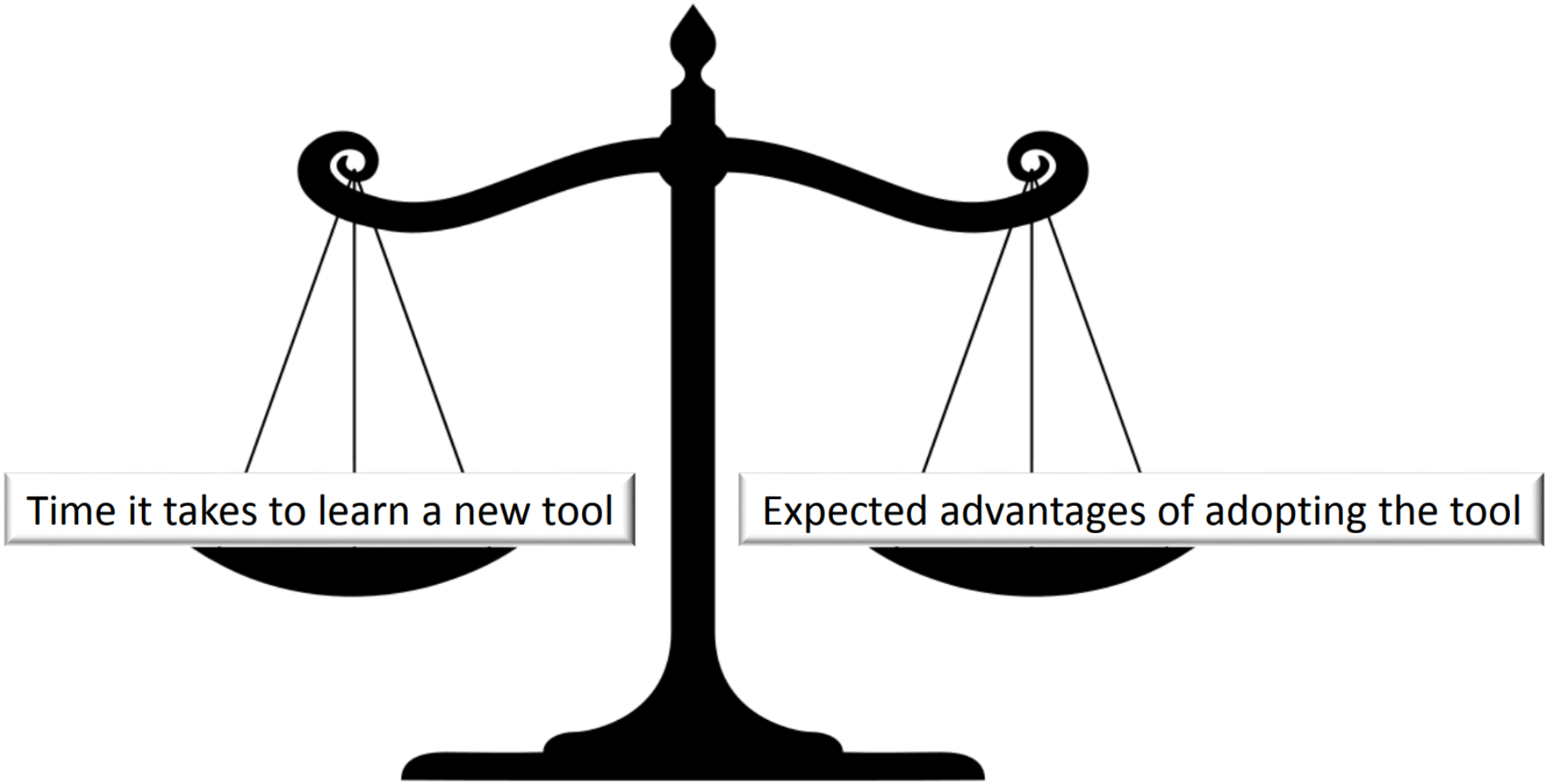
The Brick Wall

Practicing librarians may be blocked by constraining or regressive organizational structures and limitations on role responsibilities:

"I learn new skills, but I still need to do my old job."

Burton, Matt and Lyon, Liz and Erdmann, Chris and Tijerina, Bonnie (2018) *Shifting to Data Savvy: The Future of Data Science In Libraries.* Project Report. University of Pittsburgh, Pittsburgh, PA. http://d-scholarship.pitt.edu/33891/

TIME



Scale image via Wikimedia at https://commons.wikimedia.org/wiki/File:Balanced_scale_of_Justice.svg. Public Domain.

Hurdles to Implementing Data Science in Practice

Knowing that time is limited and it would take even more time to learn a new tool, we revert back to using tools that we are comfortable with.

Hurdles (Scenes from a Track Meet) photograph courtesy Phil Roeder via Flickr. Licensed under CC BY 2.0.

Jumping the Hurdles

We jump the hurdle by adopting a new tool, but we stop after one hurdle

We skip documenting because it's time intensive and we will remember

We want to get the report to administration → to the decision-making phase

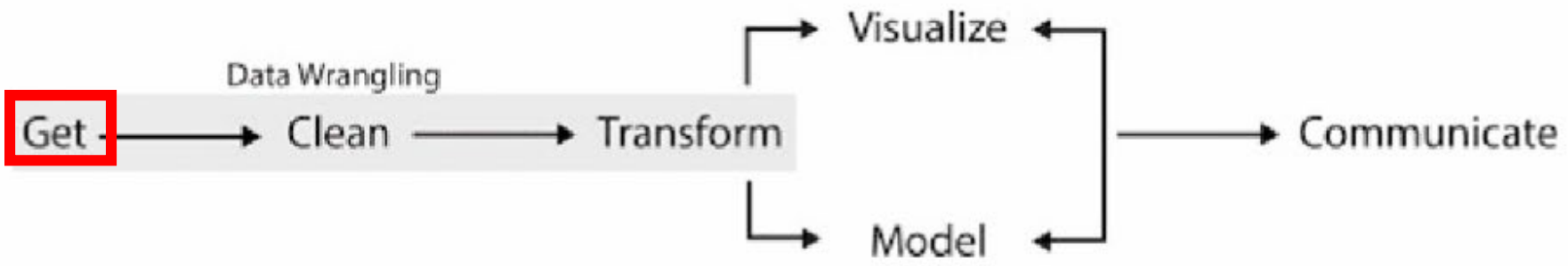
Your primary collaborator is yourself 6 months from now, and your past self doesn't answer emails.

https://dynamicecology.wordpress.com/2015/02/18/the-biggest-benefit-of-my-shift-to-r-reproducibility/

R Ainsworth – @rachaelevelyn
https://zenodo.org/record/1464853#.W8nO_mj7RPZ

https://osf.io/a5p3r/

Getting Data: Data Sources



- Directly from vendors
- ILS
- COUNTER
- Libguides

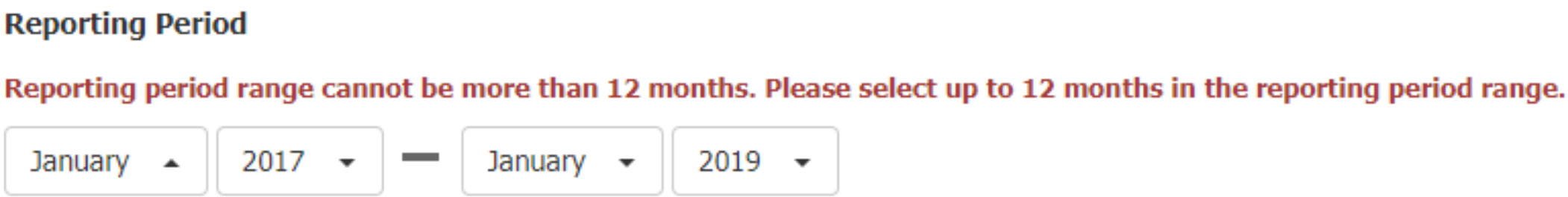
- Library staff member
- Buried somewhere on a shared departmental hard drive
- University departments
- Government & other online data portals

Platforms Limitations on Data Export

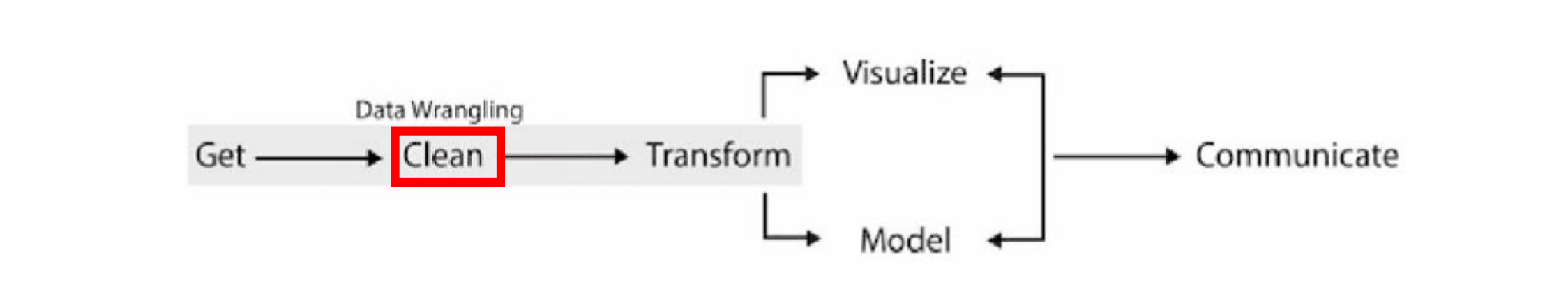
Joining Data

- Want/Need data from several reports in a single report

- Analysis usually considers multiple years of data; yet the range is limited to one year



Cleaning Data



Most visualization research assumes that input data arrive pristine, too often turning a blind eye to concerns of data formatting and quality.

Image © Sean Kandel et al., "Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data," *Information Visualization* 10, no. 4 (2011): 273.
Open Access version: http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

Messy Data

Staggered rows

```
"b11265863"^"01-23-2005"
"b11267963"^"03-25-2009"^"03-25-2009"
"b11279515"^"05-11-2004"
"b1127993x"^"07-06-2009"
"b11302793"^"08-25-2004"
"b11306920"^"07-06-2009"
"b11319732"^"02-12-2004"^"02-11-2013"
"b11324156"^"05-04-2010"
"b11340472"^"06-15-2006"
"b11344763"^"02-28-2001"
"b11348380"^"08-19-2003"
"b1135754x"^"03-28-2008"
"b11365924"^"02-24-2010"
"b11375759"^"12-07-2007"
"b11380469"^"12-13-2002"
"b11390189"^"11-03-2008"
"b11394961"^"07-31-2002"
"b11396970"^"06-12-2006"^"06-15-2006"
"b11402659"^"03-09-2006"
"b1140274x"^"02-11-2013"
```

Punctuation

Pages.Printed	User.Sessions
22,995	1,123
4,696	697
4,107	525
2,293	377

Data Errors

'1809
'1820
'1068
'1790
'1815
'19K4

Multiple Identifiers

Print ISSN	Online ISSN
0732-8303	1532-2327
1061-1983	1558-0881
0001-8732	1460-6976
0092-7872	1532-4125
1040-7782	1521-0634

No Identifiers

titles
AIS Educator Journal
Accounting Horizons
Advances in Quantitative Analysis of Finance & Accounti...
Applied Economics
Asian Review of Accounting

691026629 (pbk. : alk. paper)~69102670X (cloth : alk. paper)

https://osf.io/a5p3r/

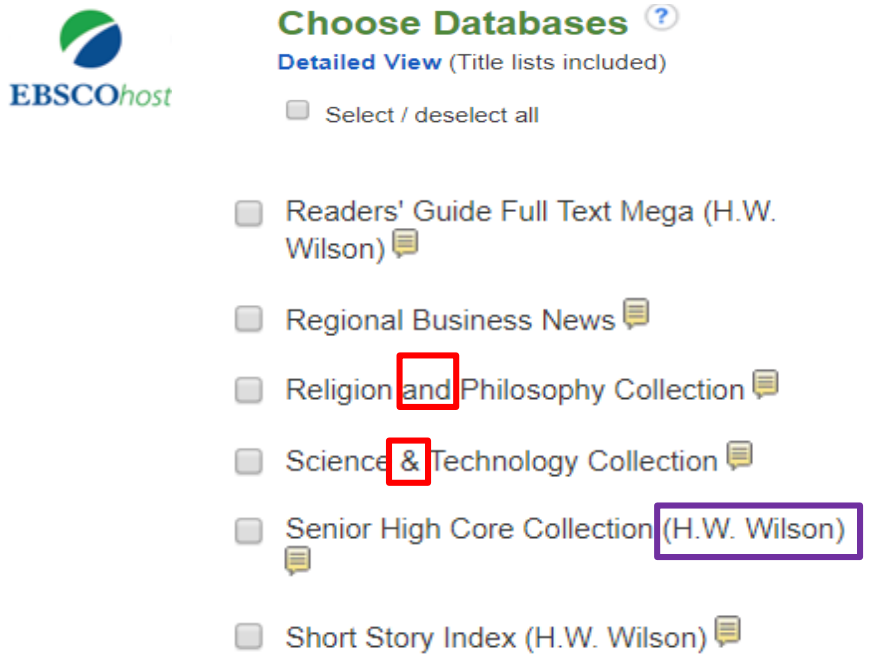
```r
ebrary_func <- function(ebks){
  # Remove commas from specified variables in the dataframe and coerce them to integers
  #
  # Arg:
  #   ebks: dataframe. Ebrary DDA past 12 month use dataset
  #
  # Returns:
  #   Same dataframe with the specified columns coerced to integer
  cols <- c("Pages.Viewed"
            , "Pages.Copied"
            , "Pages.Printed"
            , "User.Sessions"
            , "Chapter...Range.Downloads"
            , "Full.Title.Downloads")

  ebks[, cols] <- sapply(ebks[, cols], function(z) (str_replace_all(z, ",", "")))
  ebks[, cols] <- as.data.frame(sapply(ebks[, cols], as.integer))
  return(ebks)
}
```

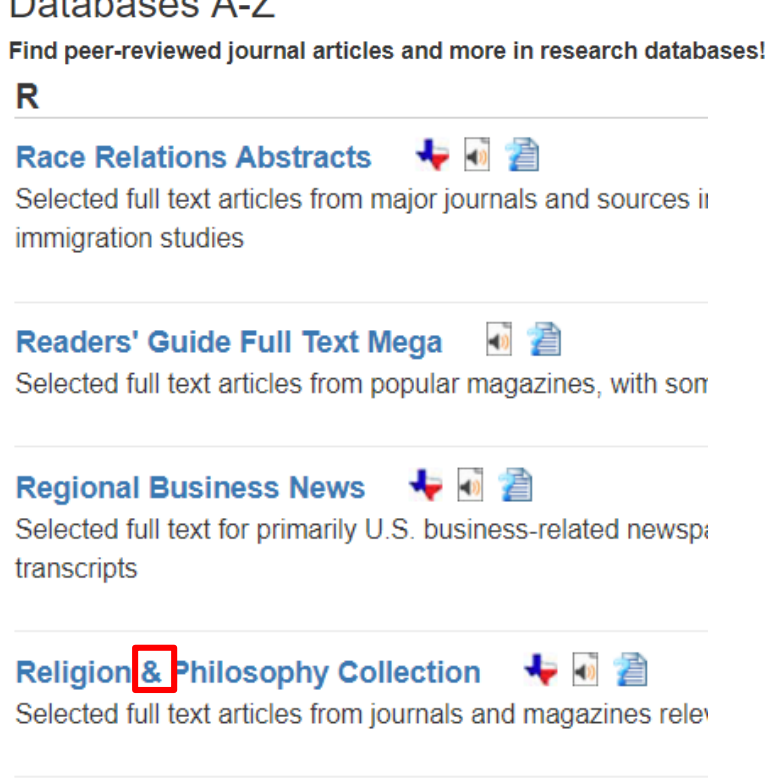
Pages.Viewed	Pages.Copied	Pages.Printed	User.Sessions
26,859	111	22,995	1,123
28,104	302	4,696	697
15,470	570	4,107	525
14,472	79	2,293	377
6,661	18	1,351	359
8,589	59	2,043	355
3,772	62	4,083	318
9,936	84	3,420	301

Pages.Viewed	Pages.Copied	Pages.Printed	User.Sessions
26859	111	22995	1123
28104	302	4696	697
15470	570	4107	525
14472	79	2293	377
6661	18	1351	359
8589	59	2043	355
3772	62	4083	318
9936	84	3420	301

Merging Non-Interoperable Data

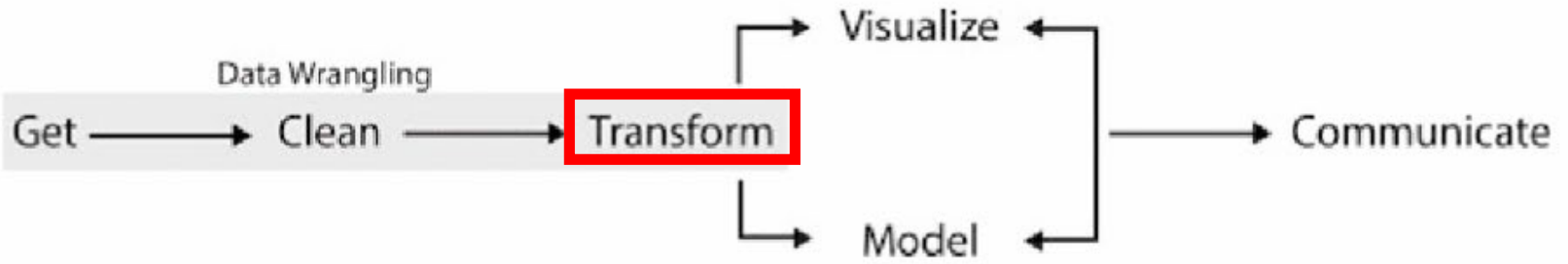
Using Text as Unique ID





https://osf.io/a5p3r/

Transforming Data: Data In Context



Data is created for a specific purpose, and is useful and meaningful within that context…

But it also almost always exists in a wider context

Transforming Data: Data In Context

- **Budget data**: library administration & planning
- **Database details data**: building the public Databases A-Z list
- **Liaison assignment data**: communicating responsibilities
- **Usage data**: collections assessment
- **Acquisitions data**: invoicing

https://osf.io/a5p3r/

Transforming Data: Data In Context

Micro: individual level

- Usage of a specific resource (journal hits, book checkouts) or database (number of searches)

Meso: group level

- Usage within a particular subject (biology), format (print books), or collection (comparison of multiple databases & usage indicators)

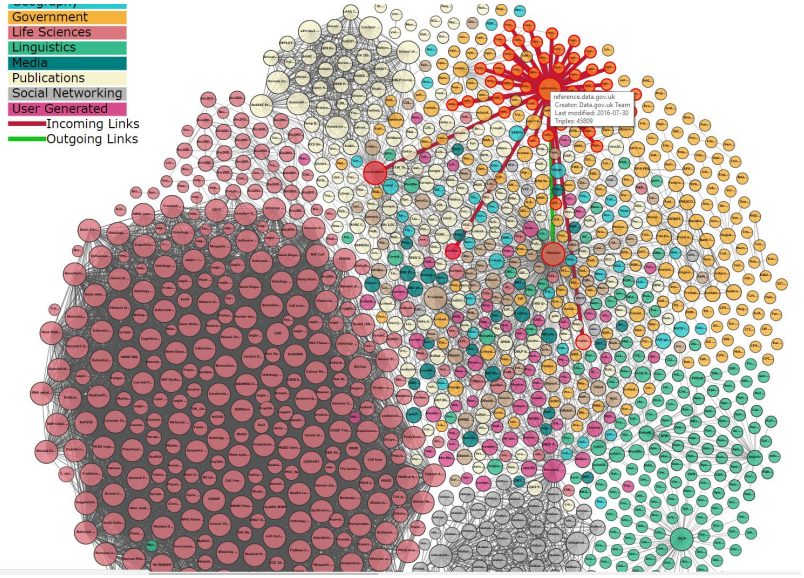
Macro: global level

- Usage of the collection in the context of the university (enrollment, student success, other dynamic indicators of collection value) or public (demographics, geography, socioeconomics, language)

https://osf.io/a5p3r/

Transforming Data: Enlarging the Context

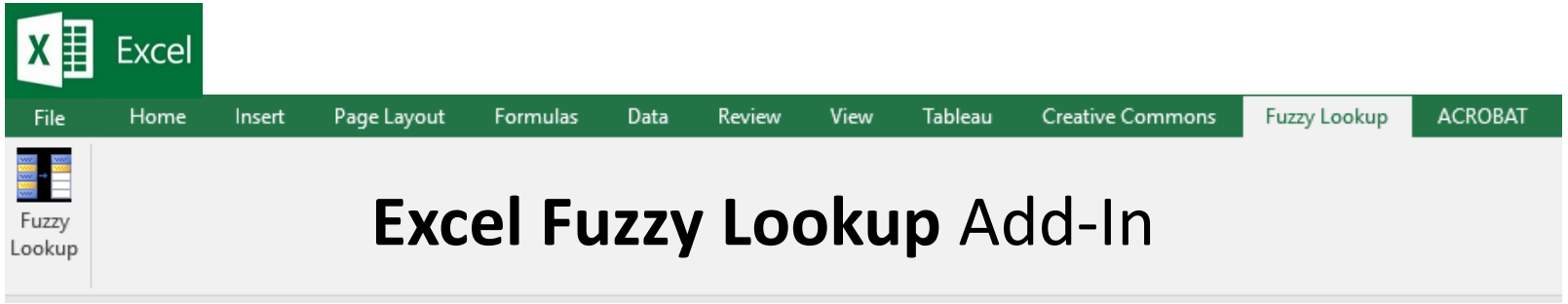
- A single dataset almost always exists within a *meso-* or *macro-* context
  - This is the whole concept behind Linked Open Data
- Visualizations of multiple datasets joined together enlarges the context
  - Time span
  - Collections
  - Vendors
  - Formats
  - Usage indicators
  - Institutional metrics



"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/" Licensed under CC BY-SA 3.0.

Transforming Data: Enlarging the Context

- Finding and cleaning **shared dimensions** between two different datasets in order to enlarge the context and join them together is a central challenge
- Matching text strings can be especially difficult
- Many important variables—such as database names—lack universal and standardized key values

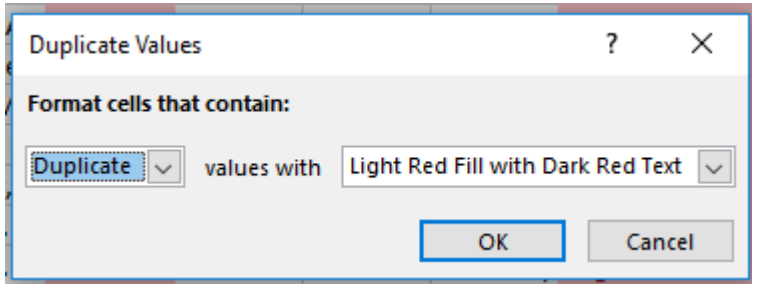


https://osf.io/a5p3r/

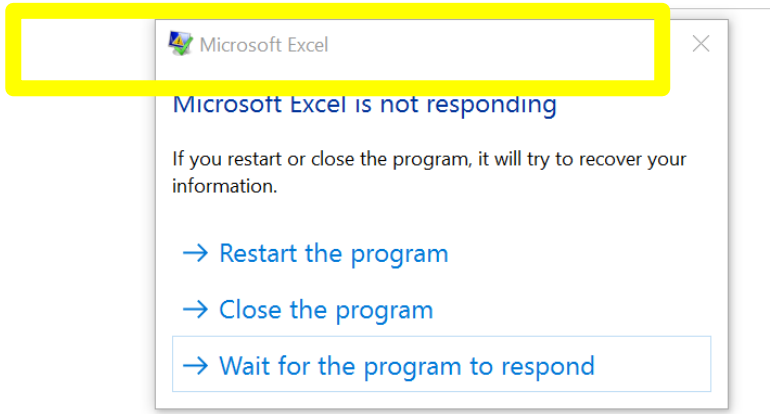
Limitations of Tools for Larger Datasets

20190212_1004775_tamucs [Read-Only] - Excel (Not Responding)

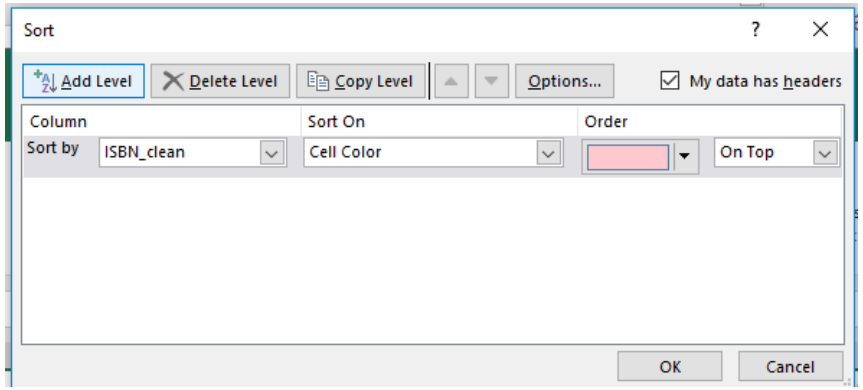
Identifying/removing duplicates



OR



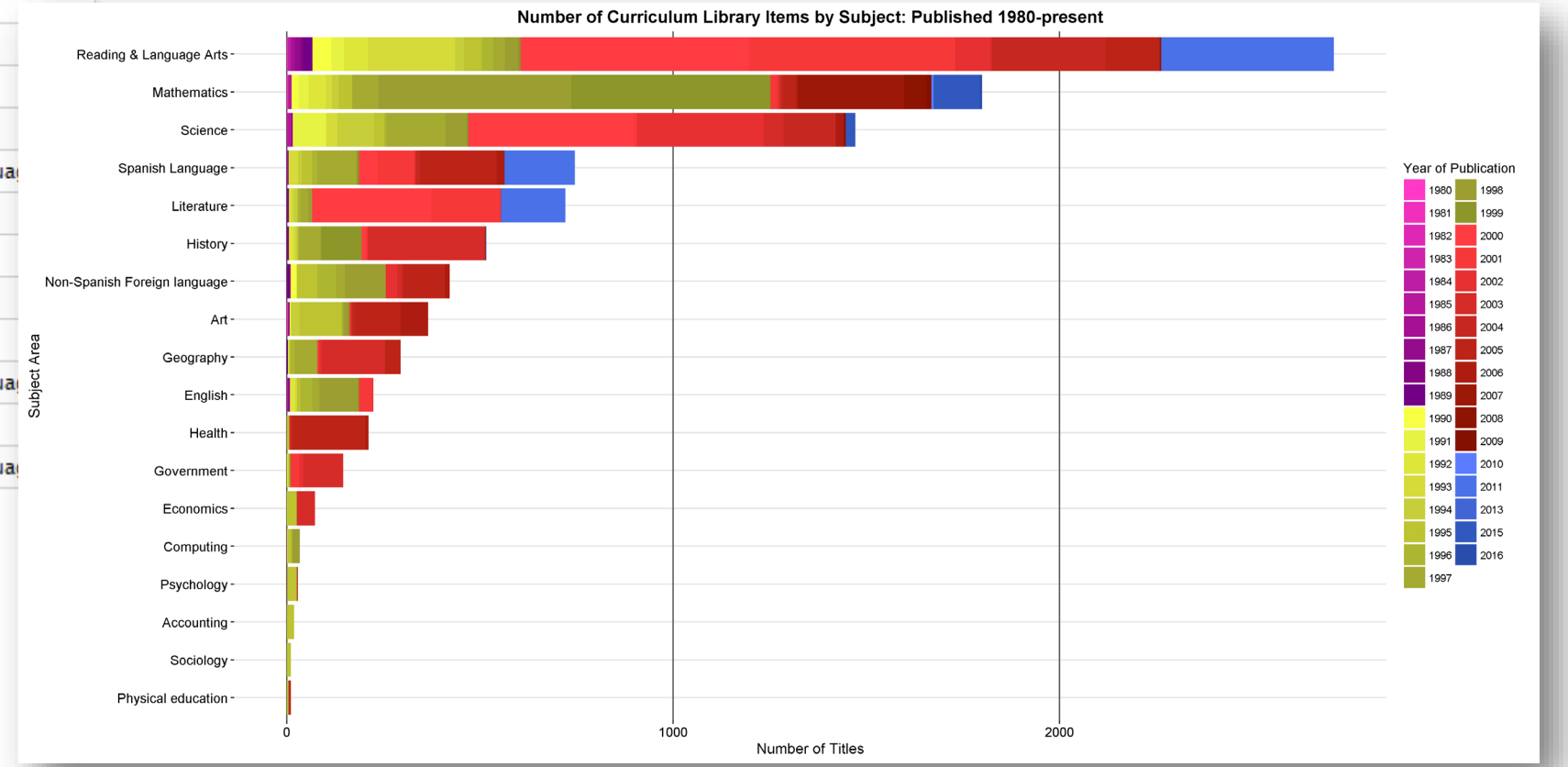
Sorting large data files



https://osf.io/a5p3r/

Recoding Variables

SUBJECT	subjRecode
Arithmetic -- Study and teaching (Elementary)~Mathe...	Mathematics
Self-perception.~Child psychology.	Psychology
School music -- Instruction and study.~Music -- Hand...	Art
Geometry -- Study and teaching (Middle school)	Mathematics
Readers -- 1950-~Reading (Elementary)~Reading (Prim...	Reading & Langua
Arithmetic.~Money -- Juvenile literature.	Mathematics
Nature study.~Ecology.	*NA*
Arithmetic -- Remedial teaching.~Arithmetic -- Problem...	Mathematics
Mathematics -- Study and teaching (Primary)~Mathem...	Mathematics
Language arts (Elementary)~Creative writing (Element...	Reading & Langua
Poetics -- Study and teaching (Elementary)	*NA*
Language arts (Elementary)~Reading (Elementary)~Si...	Reading & Langua
Nature study.~Science -- Study and teaching (Eleme...	Science



https://osf.io/a5p3r/

Tools of the Trade

Excel

MarcEdit 7

R Studio

OpenRefine

[^]*?@[^]*?\.[^]*

Regular Expressions

Image courtesy **Pietrodn** on Wikimedia Commons at https://commons.wikimedia.org/wiki/File:Email-regex.svg. Licensed under CC BY-SA 2.5.

tableau

Examples

https://osf.io/a5p3r/

Getting Journal Metrics with no ISSNs

titles
AIS Educator Journal
Accounting Horizons
Advances in Quantitative Analysis of Finance & Accounti...
Applied Economics
Asian Review of Accounting
Auditing: A Journal of Practice and Theory
Behavioral Research in Accounting

DOAJ

SerialsSolutions®

SJR Scimago Journal & Country Rank

titles.y	dist	UHCL.Subject	Rank	Title	Type	Issn	SJR	SJR.Best.Quartile	H.index	Total.Docs...2016.	Total.Docs...3years.
journal of international accounting, auditing and taxation	5	Accounting	10384	Journal of International Accounting, Auditing and Taxation	journal	ISSN 10619518	0.400	Q2	31	8	3
rand journal of economics	5	Economics	503	RAND Journal of Economics	journal	ISSN 07416261	3.340	Q1	87	39	9
journal of macroeconomics	5	Economics	6572	Journal of Macroeconomics	journal	ISSN 01640704	0.675	Q2	34	89	29
ai magazine	5	Healthcare Administration	9001	AI Magazine	journal	ISSN 07384602	0.483	Q2	59	1	9
health care manager	5	Healthcare Administration	15346	Health Care Manager	journal	ISSN 15255794, 1550512X	0.225	Q3	21	44	14
journal of applied research	5	Management	18817	Journal of Applied Research	journal	ISSN 1537064X	0.158	Q3	17	0	
communications of the acm	5	Management Information Systems	3151	Communications of the ACM	journal	ISSN 00010782	1.185	Q1	170	322	89
ieee embedded systems letters	5	Management Information Systems	11290	IEEE Embedded Systems Letters	journal	ISSN 19430663, 19430671	0.357	Q2	16	23	7
health care manager	5	Management Information Systems	15346	Health Care Manager	journal	ISSN 15255794, 1550512X	0.225	Q3	21	44	14
imaging science journal	5	Management Information Systems	16713	Imaging Science Journal	journal	ISSN 13682199	0.197	Q2	17	48	16
financial review	4	Economics	2397	Financial Review	journal	ISSN 07328516, 15406288	1.414	Q1	10	21	8

https://osf.io/a5p3r/

Paths Forward: Structured Data

Initiative for Open Citations

The Initiative for Open Citations `I4OC` is a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation data.

Crossref

unpaywall

ORCID

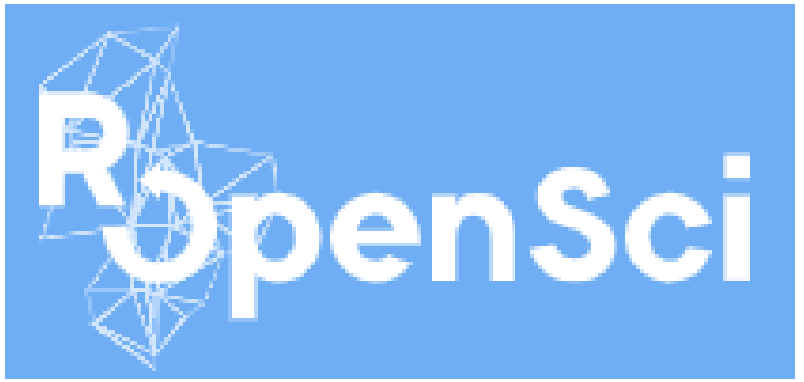
Connecting Research and Researchers

Dimensions

https://osf.io/a5p3r/

Paths Forward: rOpenSci

https://ropensci.org/packages/



bib2df: Parse a BibTeX File to a data.frame

jaod: Directory of Open Access Journals Client

rcrossref: Client for Various 'CrossRef' 'APIs'

rdpla: Client for the Digital Public Library of America ('DPLA')

roadoi: Find Free Versions of Scholarly Publications via Unpaywall

rorcid: Interface to the 'Orcid.org' 'API'

Paths Forward: Pledges

1. I will **teach** my graduate students about reproducibility.
2. All our research code (and writing) is under **version control**.
3. We will always carry out **verification and validation** (V&V reports are posted to figshare)
4. For main results in a paper, we will **share data, plotting script & figure under CC-BY**
5. We will **upload the preprint** to arXiv at the time of submission of a paper.
6. We will **release code** at the time of submission of a paper.
7. We will add a **"Reproducibility" declaration** at the end of each paper.
8. I will keep an **up-to-date web presence**.

https://osf.io/a5p3r/

"Reproducibility PI Manifesto", L. A. Barba. (13 December 2012). 10.6084/m9.figshare.104539
Presentation for a talk given at the ICERM workshop **"Reproducibility in Computational and Experimental Mathematics"**. Published on fig**share** under CC-BY.

Paths Forward: Open Science Framework

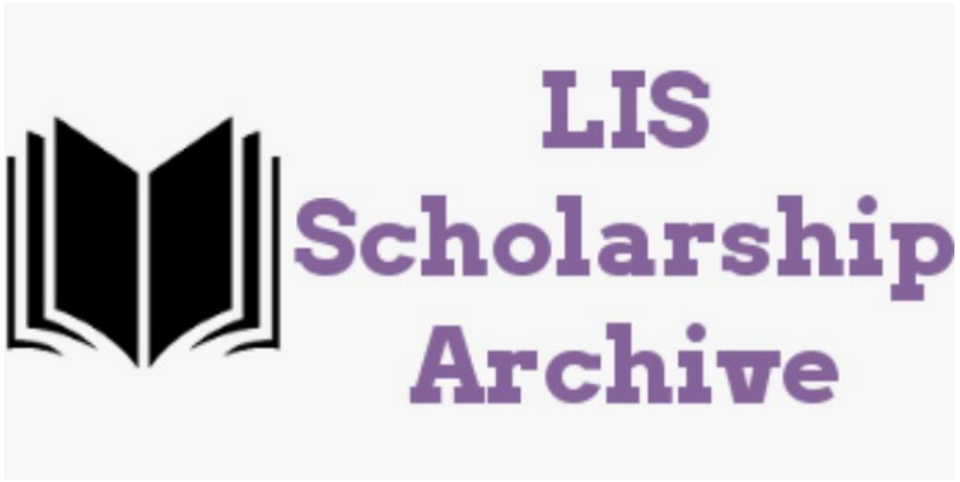
- Website: https://osf.io
- “A scholarly commons to connect the entire research cycle”
- Designed to be discipline agnostic, providing the basic tools so you can create a structure for existing workflows and processes
- Provide tools to help researchers be more efficient and effective in the research itself

Paths Forward: Open Science Framework

- A private system by default
- For managing one's own data and materials for one's own use
- So they can be more efficient in the work they do
- And provide incentives with professional benefits for making that work open for others

https://osf.io/a5p3r/

Paths Forward: Sharing Our Challenges

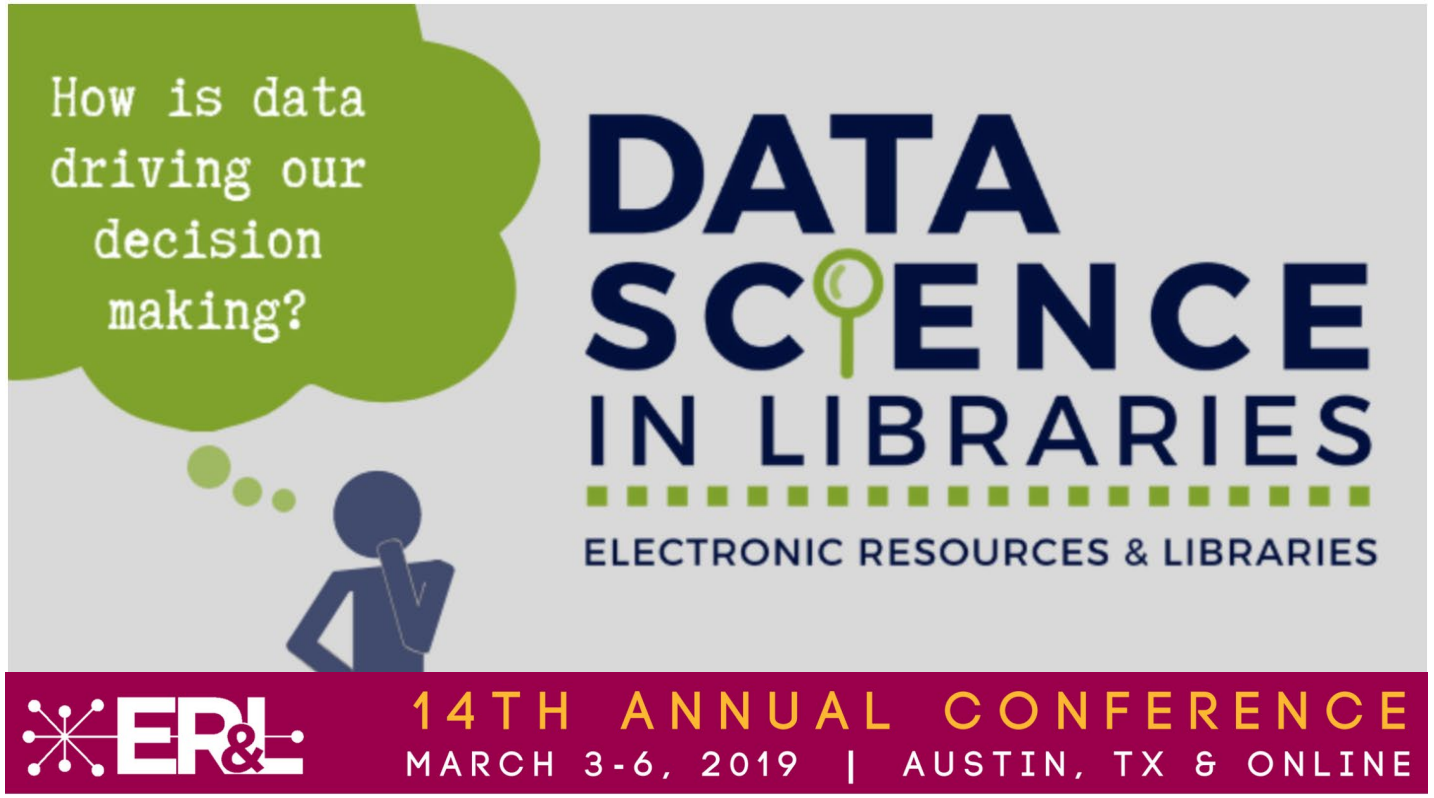


https://osf.io/preprints/lissa/

https://osf.io/a5p3r/

Paths Forward: Library Workflow Exchange

http://www.libraryworkflowexchange.org/



HATHI TRUST RESEARCH CENTER

HTRC Digging Deeper, Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

ICPSR

asis&t

Association for Information Science and Technology

iassist

International Association for Social Science Information Services & Technology

https://osf.io/a5p3r/

Shifting to Data Savvy: The Future of Data Science In Libraries

Matt Burton
Liz Lyon
Chris Erdmann
Bonnie Tijerina

https://librarycarpentry.org

Data Science in Libraries IMLS Grant

Matt Burton, Liz Lyon, Chris Erdmann, Bonnie Tijerina

http://d-scholarship.pitt.edu/33891/1/Shifting%20to%20Data%20Savvy.pdf

Suggestions & Questions

Shannon Burke, Coordinator of Monograph Acquisitions

sburke@library.tamu.edu

Clarke Iakovakis, Scholarly Services Librarian

clarke.iakovakis@okstate.edu

View data, code, and other materials at https://osf.io/a5p3r/